The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesis Titles

Lisna Zahrotun[1], Nila Hutami Putri[2], Arfiani Nur Khusna[3]
Informatics Engineering Study Program
Ahmad Dahlan University
Yogyakarta, Indonesia
lisna.zahrotun@tif.uad.ac.id[1], nila.hutami12@gmail.com[2], arfi.khusna@tif.uad.ac.id[3]

***Abstract*—One of graduation requirements at university is completing undergraduate thesis. At Industrial Engineering Universitas Ahmad Dahlan, undergraduate thesis titles are documented by thesis coordinator. The problem is that students are less knowledgeable on thesis topics, so they do not really know the previous students' thesis topics. Based on the problem, this research aims at developing a program to classify thesis title so the knowledge on the trend of thesis title topic can be got. The method used in this research was K-Means clustering, while range measurement method used was cosine similarity. The testing used Silhouette Coefficient method. The phases from text mining were tokenizing, filtering, stemming, similarity, classifying, testing. The result of this research is a program that can process the title data into trend group pattern of thesis title topic.**

From 138 data obtained, there are three clusters arranged based on the field on Industrial Engineering study program. Silhouette Coefficient testing shows score of 0.5674 that shows the clustering result is classified low. It occurs since the textual data of the thesis title is too widely distributed, so the title has relatively low similarity score

Keywords— thesis title, cosine similarity, k-means, clustering, text mining

I. INTRODUCTION

The knowledge on the trend of undergraduate students' thesis topic at university generally and study program specifically can positively give benefits for both curriculum development and roadmap planning for institution scaled research. However, technology to quickly get overall information from the result of students' thesis is really limited if compared to the available storing technology. Another problem found at university is the automatic thesis title clustering process as at Industrial Engineering Study Program Universitas Ahmad Dahlan that had never carried out thesis clustering. All this time, Industrial Engineering had not had documented data that showed the trend of thesis title at every year. Whereas, approximately 50 students of this study program graduate every year. It means there were about 50 undergraduate thesis titles produced at this study program every year. As a result the thesis title data was only there on the coordinator in excel file format and had not be published. If those thesis titles were classified, it can ease in giving reference for students in taking thesis title topic.

One of techniques used in managing document in text format is text mining. The scope of text mining involves clustering, classification, dimension reduction, topic modeling, and similarity computing [1]. Several researches related to text mining which have been conducted were similarity analysis on clustering by using shared nearest neighbor (SNN) method [2]. The use of K-Means method is also in clustering document in text format [3]. In addition, K-Means method is one of ten best data mining algorithm[4]. Another application is text mining for Arabic alphabet [5]. Review of 50 years clustering K-means [6]. and Text document clustering by using SNN method [7]. Exploring clustering methods in classifying data in text format and An experimental result from S. Jaiganesh (2015) proved that the cosine measure is the best comparability measure for the k-means calculation [8]. K means clustering algorithm is an algorithm that is sensitive to outliers [8].

From the existed problem and the ability of K means algorithm to group text data, trend clustering on thesis titles is done by using K-Means Clustering with Cosine Similarity. It is expected that this research can classify the thesis title into trend group pattern of thesis topic as consideration in giving reference to students and curriculum development.

II. RESEARCH METHOD

A. Research Data

Research data used in this research was data of undergraduate thesis title of Industrial engineering study program with the data number of 138.

B. Text Mining

Text mining involved traditional data mining algorithm such as clustering, classification. Text mining is a repeating process involving analysis repetition by using different setting and using or excluding particular requirements for a better result. The result of this step can be form of document collection, long-term or multi-terms topic list, or rules to solve classification problem. The steps of text mining) is shown on Fig. 1 [9]

978-1-5386-6940-2/18/$31.00 ©2018 IEEE

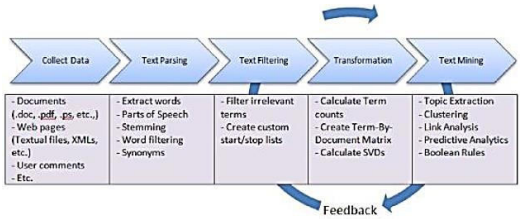


Fig. 1. *Text Mining Process Flow*

1. Collect Data

The first step in text mining research project is collecting textual data that is needed to explore the quality information from the textual data.

2. Text Parsing

Textual data parsing is started by taking character sequence (such as sentence sequence in text document) and break it down into tokens (units, where one unit exists in one word, number, or punctuation). This process is called as tokenization. Tokenization process is the process of breaking up sentence or document into word pieces [10]. After the tokens are found in the document, normalization is done to eliminate the word complexity or called as stemming. For computation, stemming functions to reduce all similar word variation.

3. Text Filtering

Corpus in thousand documents probably contains a lot of irrelevance, both to differentiate one document and another and to summarize document. Tracking particular terms manually to delete irrelevant terms is a work that often takes time and a subjective task from all text mining steps.

4. Calculate Terms Count

Terms counting is carried out by calculating *Term Frequency-Inverse Document Frequency* (tf-idf) [11]. Tf-idf is a numerical statistic that reflects how important a word is to a document in a collection. Tf - IDF is often used as weighting factor in searches of information retrieval and text mining. State that word/term count is explained with equation (1) [12]

$$\text{tfidf}\,(t_i, d_j) = \text{tf}\,(t_i, d_j) \times \log\left(\frac{N}{N\,(t_i)}\right) \quad (1)$$

where :

$\text{tfidf}\,(t_i,d_j)$ = word/ term count towards document d_j
$\text{tf}\,(t_i,d_j)$ = number of times word/ term t_i appears in the document d_j
N = total number of documents
$N(t_i)$ = number of document with word/ term t_i in it.

5. Cosine Similarity

Cosine similarity is a method to measure similarity between two vectors by measuring the cosine of the angle between them. The cosine of 0 is 1, and the bigger angle; the smaller similarity [13]. The cosine similarity is better than jaccard similarity and combination between cosine and jaccard similarity [14]. Cosine similarity is shown in equation (2).

$$\text{sim}(X_a, X_b) = \cos\theta = \frac{X_a . X_b}{\|X_a\|\|X_b\|} = \frac{\sum_{i=1}^{d} X_a^i . X_b^i}{\sqrt{\sum_{i=1}^{d} (X_a^i)^2} \;\; x \;\; \sqrt{\sum_{i=1}^{d} (X_b^i)^2}} \quad (2)$$

Where :

X_a = number of terms contained in the document a
X_b = the number of terms contained in the document b
i = 1, namely the number of terms in each document
d = each term multiplied by the number of terms in the document.

6. K-Means Clustering

K-means clustering is a method that includes in partitioning clustering algorithm group. K-means clustering algorithm is a centroid based methodology in wich the similarity of group is measured in concern to the mean estimations of the documents [8]. K-means clustering algorithm is as following:

a) Determining the number of cluster
b) Determining the center of each cluster
c) Calculating centroid or average of the data in each cluster and classify each data to centroid or the average based on the closest range
d) Back to step c if there is still data that migrate to another cluster or if the centroid score changes.

7. Silhouette Coefficient Testing

Silhouette coefficient is used to find out the cluster quality and strength, how good an object is placed in a cluster [13]. Score of silhouette coefficient is shown in equation (3).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

where :

s(i) = value of silhouette validity with object i
a (i) = average distance between objects i with all objects in the same cluster (intracluster)
b (i) = average distance between objects i with all objects in the nearest cluster (nearest cluster)
max = maximum

III. Result and Discussion

In this research, data obtained from Industrial engineering study program is processed through text mining phases:

A. Collect Data

Collect data is carried out by collecting data. The data was obtained from coordinator of undergraduate thesis of Industrial Engineering study program. The data obtained is saved in excel format. To process the document, the data is inputted to system and stored in database by using MongoDB. The process of Load Data is shown on Fig. 2.

	0	1	2	3
0	no	nim	nama	judul
1	1	8019001	AGUS DWI KRISTIYANTO	Analisis Pengaruh Kualitas Produk dan Kualitas Pelayanan Terhadap Kepuasan Pelanggan Menggunakan Analisis Regresi Linier Berganda (Studi Kasus Pada Cafe Semesta Kota Baru Yogyakarta
2	2	8019002	ARIEF WICAKSONO	Perancangan Alat Press Produk Placemat Berbahan Enceng Gondok dengan Pendekatan Ergonomi Guna Mengurangi Kecacatan Produk dan Meningkatkan Produktivitas
3	3	8019003	TRI NOVI ANDRIANTORO	Perancangan Ulang Alat Pengepres Sandal Dengan Pendekatan Ergonomi Untuk Memperbaiki Posisi Kerja Meningkatkan Output, Dan Mengurangi Tingkat Kecacatan Produk
4	4	8019005	ALFAHMI	Perancangan Ulang Alat Press Sol Sandal Kulit Dengan Pendekatan Ergonomi Untuk Meningkatkan Produktivitas Pada Stasiun Pengepresan (Studi Kasus di UKM Marlan's, Jln. Nyi Adisoro I, Depokan, Yogyakarta
5	5	8019006	DIAN RIYADI	Pengendalian Kualitas dalam Upaya Menekan Tingkat Kecacatan Produksi Penyamakan Kulit Dengan Menggunakan Alat Bantu Statistik
6	6	8019008	YAHYA HASAN	Analisis Faktor-Faktor yang Mempengaruhi Kepuasan Konsumen di Pamela Futsal
7	7	8019009	DITO AULIA DHARMAWAN GUFRON	Faktor Penghambat dan faktor Pendorong yang Mempengaruhi Niat Beli Pelanggan Belanja Online pada Situs Jual Beli Online dengan Menggunakan Metode Structural Equation Modelling (SEM)

Fig. 2. The Process of Load Data

In this process, data selection per year is also carried out. The document of title per year is shown on Fig. 3.

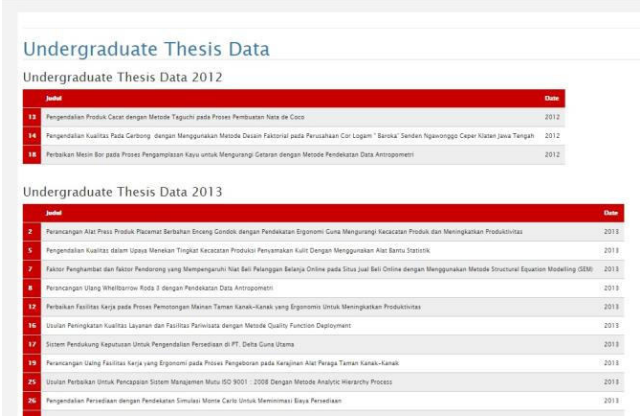
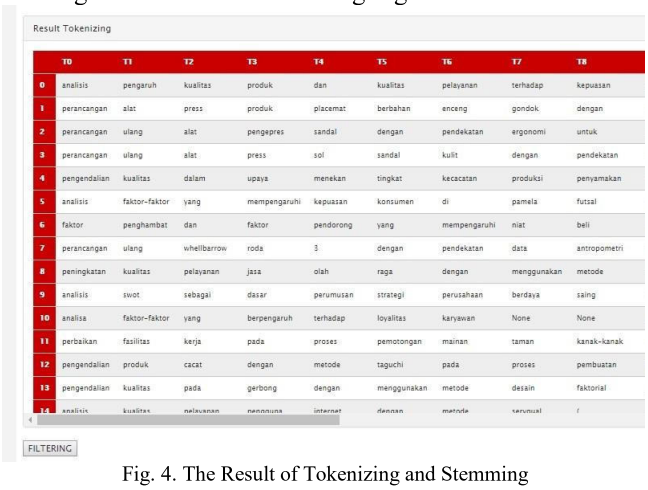


Fig. 3. The Result of Selection Process

B. *Text Parsing*

The process of text parsing is divided into two which are tokenizing and stemming. Tokenizing is used to break down the sentence into words. After tokenizing, stemming and filtering is done. The result of tokenizing, stemming and filtering is shown as the following Fig. 4.



	T0	T1	T2	T3	T4	T5	T6	T7	T8
0	analisis	pengaruh	kualitas	produk	dan	kualitas	pelayanan	terhadap	kepuasan
1	perancangan	alat	press	produk	placemat	berbahan	enceng	gondok	dengan
2	perancangan	ulang	alat	pengepres	sandal	dengan	pendekatan	ergonomi	untuk
3	perancangan	ulang	alat	press	sol	sandal	kulit	dengan	pendekatan
4	pengendalian	kualitas	dalam	upaya	menekan	tingkat	kecacatan	produksi	penyamakan
5	analisis	faktor-faktor	yang	mempengaruhi	kepuasan	konsumen	di	pamela	futsal
6	faktor	penghambat	dan	faktor	pendorong	yang	mempengaruhi	niat	beli
7	perancangan	ulang	whellbarrow	roda	3	dengan	pendekatan	data	antropometri
8	peningkatan	kualitas	pelayanan	jasa	olah	raga	dengan	menggunakan	metode
9	analisis	swot	sebagai	dasar	perumusan	strategi	perusahaan	berdaya	saing
10	analisa	faktor-faktor	yang	berpengaruh	terhadap	loyalitas	karyawan	None	None
11	perbaikan	fasilitas	kerja	pada	proses	pemotongan	mainan	taman	kanak-kanak
12	pengendalian	produk	cacat	dengan	metode	taguchi	pada	proses	pembuatan
13	pengendalian	kualitas	pada	gerbong	dengan	menggunakan	metode	desain	faktorial
14	analisis	kualitas	pelayanan	pengguna	internet	dengan	metode	servqual	r

Fig. 4. The Result of Tokenizing and Stemming

C. *The process of Document Clustering*

Before clustering, document from preprocessing result will be tested with TF IDF. After that, clustering by using K-Means method is carried out. In the clustering process, 3 clusters are used which are Product Ergonomics and Design, Quality Management, and Production system. The result of clustering is shown on the following Fig. 5.

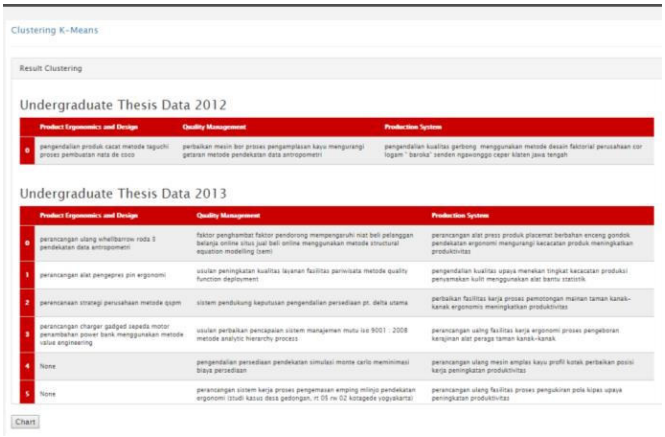


Fig. 5. Clustering Result

Clustering result is then presented in the form of undergraduate thesis topic trend graphic. The graphic is shown on the following Fig. 6.

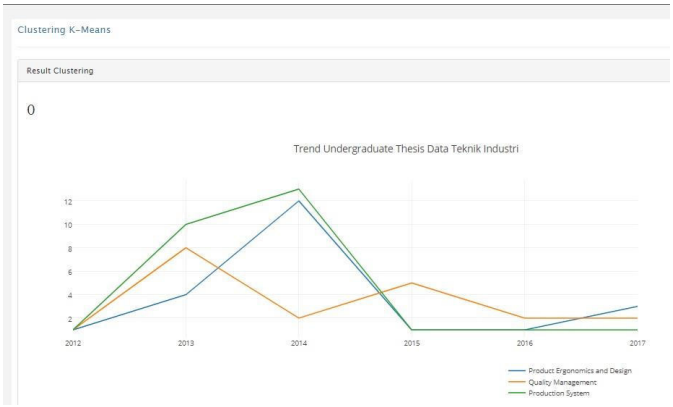


Fig. 6. Undergraduate Thesis Topic Trend Graphic

D. *Testing*

From 138 data clustering by using K-Means method, the result of Silhouette Coefficient is 0.5674. With the number of cluster of 3, which are Product Ergonomics and Design, Quality Management, and Production system, the result is relatively low. From the analysis, the wide distribution of thesis title makes the testing score become low. The low score of silhouette coefficient is because the data used is thesis titles written in different words from one to another title.

Clustering result with Silhouette Coefficient 0.5674 can be seen on Fig. 6. And from the graphic on figure 6, it can be seen that undergraduate thesis topic trend of industrial engineering from the three fields is relatively stable but has some difference as following:

1. In 2012-2013 topic on Product Ergonomics and Design and Quality Management increased, but topic on Production System decreased.
2. In 2015-2016 topic on Product Ergonomics and Design and Quality Management decreased, but topic on Production System increased.

IV. CONCLUSION

Based on the result of the research on the Implementation Text Mining on Undergraduate Thesis Title

Topic Trend Clustering at Industrial Engineering Study Program Universitas Ahmad Dahlan by Using K-Means Clustering with Cosine Similarity", it can be concluded that:

1. A program with python programming language that can group the thesis titles to show the trend every year has been developed.
2. From the silhouette coefficient testing, the result of silhouette coefficient is 0.5674.

The low score of silhouette coefficient is because the data used is thesis titles written in different words from one to another title. In addition, textual data of thesis title is too widely distributed so it results in low similarity range score from one to another with condition of range cluster = 3. However, if condition of range cluster = 9, 10, 11, the score of silhouette coefficient is relatively high which is 1. Because of the more number of cluster, the textual data of thesis title will position based on the members which have high similarity score.

ACKNOWLEDGMENT

This research has been supported by internal research grant of Universitas Ahmad Dahlan with Fundamental research scheme number PF-065/SP3/LPPM-UAD/VI/2018 date 9 June 2018

REFERENCES

[1] V. B. Kobayashi, S. T. Mol, H. Berkers, G. Kismihok, and D. N. Den Hartog, "Text Mining in Organizational Research," *Organ. Res. Methods*, vol. 21, no. 3, pp. 733–765, 2018.

[2] A. K. Patidar, J. Agrawal, and N. Mishra, "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach," *Int. J. Comput. Appl.*, vol. 40, no. 16, pp. 1–5, 2012.

[3] W. Junjie, *Advances in K- Means Clustering : a Data Mining Thinking*. Springer Science & Business Media, 2012.

[4] W. Xilon and K. Isua, *The Top Ten Algorithms in Data Mining.* London: CRC Press Tailor & Francis Group, 2009.

[5] S. A. Salloum, A. Q. Alhamad, M. Al-emran, and K. Shaalan, "A Survey of Arabic Text Mining," *Sringer Int. Publ.*, pp. 417–431, 2018.

[6] A. K. Jain, "Data clustering : 50 years beyond K-means," *Pattern Recognit. Lett.*, pp. 1–16, 2009.

[7] L. Zahrotun, "Text Mining for Internship Titles Clustering Using Shared Nearest-Neighbor Method," *Comput. Eng. Appl.*, vol. 6, no. 3, 2017.

[8] N. Garg and R. K. Gupta, "Exploration of Various Clustering Algorithms for Text Mining," *Int. J. Educ. Manag. Eng.*, vol. 4, no. July, pp. 10–18, 2018.

[9] G. Chakraborty, M. Pagolu, and S. Garla, *PREVIEW: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. 2013.

[10] C. D. Manning, P. Raghavan, and H. Schutze, *introduction to Information Retrieval*. Camridge: Cambridge University Press, 2008.

[11] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[12] N. L. Processing and T. Mining, *Natural Language Processing and Text Mining*. USA: Springer, 2006.

[13] C. Plattel, "Distributed and Incremental Clustering using Shared Nearest Neighbours," Utrecht University, 2014.

[14] L. Zahrotun, "Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," vol. 5, no. 1, pp. 11–18, 2016.